Open Data Lab Annual Report - 2018

The Open Data Lab Collaboration

January 2, 2019

The Team

Pete Alonzi came to data science by way of the particle physics community. As a result he has great interest in making data open and usable to broad audiences. He sereves as a data scientist at the DSI and is the project manager for the Open Data Lab.

Phil Bourne "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Tim Clark "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Daniel Mietchen "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Lane Rasberry "At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excuri sint occaecati cupiditate non provident,At vero eos et accusamus et dent, "

Letter from the editor

hello world

Contents

Chapter 1

Overview

1.1 What is the Open Data Lab?

OPEN

We encourage all users to be as open as possible with every aspect of their work. That may be in opening up their data sets, publication, source code, or ...

DATA

We take an expansive definition of data. Everything from traditional data, to code, to workflows, to published material, and so on is considered data to us. We provide a place for all things digital data.

LAB

We provide a place where the power of contemporary computing can be brought to bear against data resources. Given the scale of data today this means colocating the data and computational resources.

Open Data Lab

The Open Data Lab is a resource to provide state of the art computing and data infrastructure to researchers, students, and sharers. It is guided by the principles of science and openness.

1.2 User Archetypes

There are many potential use cases for the Open Data Lab. In this section we describe the three cases that have been tested so far. They are: the Collaborator, someone who is working on a research project; the Student, someone who is using the Open Data Lab to learn about data science; and the Sharer, a person with data who wants to open it up to a broader audience.

The Collaborator

This archetypal person uses the Open Data Lab to conduct research. They access data and computational resources that are colocated. This colocation facilitates lower latency and increased performance. A wide range of services can be provided globally by AWS and locally through UVA HPC resources. Sample workflow:

1. Request a user account on the Open Data Lab
2. Once per collaboration:
   (a) Load data
   (b) Provision computational resources
3. Conduct research operations
4. Register resulting products in Dataverse

The Student

This archetype uses the Open Data Lab to facilitate learning. An example would be someone who participates in a workshop where and ODL notebook instance powered by AWS SageMaker provides the working environment. Sample workflow:

1. Request a user account on the Open Data Lab
2. Logon to AWS console to launch Jupyter
3. Use Jupyter during the workshop

The Sharer

This archetype is a user who owns data and wants to make it available. There are many mechanisms for sharing the data ranging from RESTful API of S3, to a SPARQL endpoint. Sample workflow:

1. Request a user account on the Open Data Lab
2. Load data into an S3 bucket
3. Configure one of the following
   (a) SPARQL endpoint
   (b) API Gateway to access S3
   (c) S3 permissions for a SageMaker notebook
   (d) ...

1.3 User Summary

group	projectID	# members	type
Bourne-Mura	bamc	4	MSDS Capstone
CBW	cbwc	4	MSDS Capstone
Wiki	wiki	9	MSDS Capstone
Mental Health	miip	6	SYS Capstone
Women Terror Recruitment	watr	2	Presidential Fellow
Healthy Markets	hmtt	5	DSI Research
Independent Study	pmis	1	DSI Research
Linked Open Data	nept	2	External Data
Spark	sprk	17	Education
GitHub	gith	9	Education
ORCI	orci	2	ODL Development
ML under	mlunder	7	Club
ML grad	mlgrad	3	Club
Rivanna	–	11	Local
Ivy	–	6	Local
ODL-education	–	26	Education Users
ODL-users	–	68	Unique Users

1.4 A phased approach

The first three phases of the Open Data Lab have been outlined. Phase 0 focused on pre investigation and decided on what technology to test in Phase 1, the closed beta. Phase 2 is an open Beta and will serve the community of Charlottesville and other associated research and educational efforts.

Phase	type	start	end
0	alpha	FEB 2018	JUN 2018
1	closed beta	JUL 2018	MAY 2019
2	open beta	TBD	–

Short Term Goals

1. Explore technology to solve open Open Data Lab questions
2. Serve Data Science Institute users
3. Understand user archetypes

Long Term Goal

Change the way science is done.

1.5 What's next for the Open Data Lab?

The year 2019 will be a big year for the Open Data Lab. The gathering of information and skill from the closed beta has gone well. While continuing the closed beta through the spring semester 2019 the Open Data Lab will be seeking to add staff to prepare for the open beta launch.

Chapter 2

Key Developments

2.1 Phase 1 - Closed β

2.2 Establishment of User base

2.3 Technology exploration

2.3.1 Amazon Web Services

S3

EC2

SageMaker (Jupyter)

IAM

API Gateway

Architecture Diagrams

lambda

13 TB Data Transfer

Usage Report

Support Plan

`https://aws.amazon.com/premiumsupport/compare-plans/`

2.3.2 Local UVA - Rivanna and Ivy

2.3.3 Github

2.3.4 Dataverse

2.3.5 Spark

2.3.6 SPARQL Endpoint

Chapter 3

Datasets

3.1 Healthy Markets

3.2 Numismatic

Chapter 4

Research

4.1 Bourne/Mura Capstone

4.2 DSI Wiki

Chapter 5

Education

5.1 Spark Workshop

powered by aws sagemaker

5.2 Github Workshop

taught in github itself

Chapter 6

Vital Metrics

6.1 AWS usage

6.2 FTE analysis

6.3 Budget

6.3.1 Funding

funded by the Data Science Institute healthy markets project funded off-budget for ODL